Homework #1

See Canvas for HW #1 assignment due date. Complete all of the following problems. Ideally, the theoretical problems should be answered in a Markdown cell directly underneath the question. If you don't know LaTex/Markdown, you may submit separate handwritten solutions to the theoretical problems, but please see the class scanning policy. Please do not turn in messy work. Computational problems should be completed in this notebook (using the R kernel). Computational questions may require code, plots, analysis, interpretation, etc. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

A. Theoretical Problems

Problem A.1

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known, and we are interested in an estimator for $\theta = \mu^2$. (Note that we will use the following calculuations to make comparisons to the parametric boostrap method explored below).

(a) Find the maximum likelihood estimator (MLE) for θ, denoted $\hat{\theta}$.

This will be easy if you use one of the important properties of MLEs!

Since $\bar{X}$ is the MLE for μ, $\bar{X}^2$ is the MLE of $\theta = \mu^2$ (by the invariance property of MLEs).

(b) Compute the bias of $\hat{\theta}$, denoted $Bias(\hat{\theta})$. Recall that $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.

Note that $Var(\bar{X}) = E(\bar{X}^2) - \left[E(\bar{X})\right]^2 \implies E(\bar{X}^2) = Var(\bar{X}) + \left[E(\bar{X})\right]^2 = \frac{\sigma^2}{n} + \mu^2$. Thus,

$$Bias(\hat{\theta}) = Bias(\bar{X}^2) = E(\bar{X}^2) - \mu^2 = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}$$

(c) (STAT 5010 Only) Compute the variance of $\hat{\theta}$, denoted $Var(\hat{\theta})$. (HINT: You might use a moment generating function at some point in your answer.)

$Var(\hat{\theta}) = Var(\bar{X}^2) = E(\bar{X}^4) - \left[E(\bar{X}^2)\right]^2 = E(\bar{X}^4) - \left(\frac{\sigma^2}{n} + \mu^2\right)^2 = E(\bar{X}^4) - \frac{\sigma^4}{n^2} - 2\frac{\sigma^2}{n}\mu^2 - \mu^4$. We need a way of computing $E(\bar{X}^4)$. Luckily, we know that, for *any* random variable X,

$$E(X^k) = \frac{d^k M(t)}{dt^k}\Big|_{t=0},$$

where $M(t)$ is the moment generating function for X. The moment generating function for $\bar{X}$ is

$$M_{\bar{X}}(t) = \exp\left\{\mu t + \frac{\sigma^2}{2n}t^2\right\}.$$

So, we just need to find the fourth derivative of this and plug in $t = 0$.

[Math Processing Error]

Thus,

$$Var(\bar{X}^2) = E(\bar{X}^4) - \frac{\sigma^4}{n^2} - 2\frac{\sigma^2}{n}\mu^2 - \mu^4 = \mu^4 + \frac{6\sigma^2}{n}\mu^2 + \frac{3\sigma^4}{n^2} - \frac{\sigma^4}{n^2} - 2\frac{\sigma^2}{n}\mu^2 - \mu^4 \qquad = \boxed{4\frac{\sigma^2}{n}\mu^2 + 2\frac{\sigma^4}{n^2}}$$

(d) Write down the boostrap estimators of $Bias(\hat{\theta})$ and $Var(\hat{\theta})$.

@@math2@

$$Var(\bar{X}^2) \approx 4\frac{S^2}{n}\bar{X}^2 + 2\frac{S^4}{n^2}$$

or, for the booststrap estimators (STAT 4010):

\begin{align*} \widehat{B}(\widehat{\theta}) \approx \frac{1}{B}\sum^B_{j=1}\widehat{\theta}^_j - *bar{x}*^2. *end{align*}

and

\begin{align*} \widehat{\text{Var}}(\widehat{\theta}) \approx \frac{1}{B-1}\sum^B_{j=1}\big(\widehat{\theta}^_j - *bar{\theta}* *big)*^2, *end{align*} where \begin{align*} \bar{\widehat{\theta}} = \frac{1}{B}\sum^B_{j=1}\widehat{\theta}^_j . *end{align*}

Problem A.2

Provide a brief explanation of the pros and cons of using the bootstrap for calculating confidence intervals.

The bootstrap is flexible in that it requires few assumptions (e.g., it doesn't require assumptions about the population model or the applicability of the central limit theorem). So, we can use the bootstrap in a wide range of situations (for a wide range of parameters).

The bootstrap can be computationally expensive, and some bootstrap confidence intervals may not have great coverage properties in certain situations. We've seen that there are different types of bootstrap confidence intervals, and which is best may be contextual.

B. Computational Problems

Problem B.1

Suppose that $X_1, \ldots, X_8 \overset{iid}{\sim} \Gamma(\alpha, \beta)$. Let's use the bootstrap to compute a 90% confidence interval for the population standard deviation: $sd(X) = \sqrt{\alpha/\beta^2} = \theta$.

Note:

The convention in this course will be to interpret $\Gamma(\alpha, \beta)$ as the "shape/rate" parameterization: shape = α, rate = β. But R uses the "shape/scale" parameterization: shape = α, scale = $\theta = 1/\beta$.

To be sure that you are properly simulating from the right gamma distribution, see the help file for rgamma() (run: ? rgamma). Also, see here for more information on the gamma distribution.

(a) State why a χ^2 confidence interval is not valid in this context.

You should reply on knowledge from your prereq class!

The χ^2 confidence interval is not valid because it requires a normal population. That is,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

only when $X_1, \ldots, X_n$ are normal.

(b) Generate a sample of size $n = 8$ from $\Gamma(\alpha = 3, \beta = 4)$ and calculate the true population standard deviation (in this example, we are generating data so that we can see how well our estimation procedure will do).

```
set.seed(919)
n = 8; alpha = 3; beta = 4; x = rgamma(n,alpha,rate = beta); sd(x)
sdx = sqrt(alpha/beta^2);
cat("The population standard deviation is ", sdx,".")
```

(c) Generate $B = 200$ bootstrap samples from the above sample. Print the dimension, and articulate what each row/column represents. Avoid loops! (HINT: use the replicate() function.)

```
B = 200
bootsamples = replicate(B, sample(x, n, replace = TRUE))
 dim(bootsamples)
```

Each column is bootstrap sample from our original sample.

(d) Calculate and print the sample standard deviation, s. Then, calculate s for each bootstrap sample. Denote this as s_i^*, for $i = 1, \ldots, B$. Avoid loops! (HINT: use the apply() function.) Display a histogram of the distribution of s_i^*, $i = 1, \ldots B$.

```
s = sd(x); cat("s is ", s)
sboot = apply(bootsamples, 2, sd); hist(sboot);
```

(e) Use the quantile() function to find the 5th and 95th percentile of the distribution of s_i^*. Use these values to calculate the 90% boostrap pivot confidence interval and bootstrap percentile confidence interval for θ.

```
q = as.numeric(quantile(sboot, c(0.05,0.95)));
CI = c(2*sqrt(n/(n-1))*sd(x) - q[2], 2*sqrt(n/(n-1))*sd(x) - q[1]); CI
```

(f) Interpret this confidence interval.

If we repeatedly sample from the population and use this procedure to calculate a confidence interval, approximately 90% of those intervals will cover the true parameter.

Problem B.2

Thus far, we've been looking at the *nonparametric bootstrap*. In this problem, we look at the *parametric bootstrap* as a way of estimating the bias and variance of an estimator $\hat{\theta} = \bar{X}^2$ of $\theta = \mu^2$ (in problem A.1 you calculated these values exactly).

(a) Generate $X_1, \ldots, X_{20} \overset{iid}{\sim} N(\mu = 2, \sigma^2 = 1)$, and then forget that you know μ and σ^2. Find the sample mean and sample variance.

```
set.seed(232)
n = 20; x = rnorm(n, 2, 1); xbar = mean(x); xvar = var(x);
cat("The sample mean is", xbar,". The sample variance is ",xvar,".")
```

(b) Define $\widehat{N}$ to be the distribution of the variable X_i in the population with the sample estimates plugged in for the unknown population parameters. Write down $\widehat{N}$ based on the data generated in (a).

$\widehat{N} \approx N(2.35, 0.72).$

(c) Draw $B = 500$ parametric bootstrap samples from $\widehat{N}$, and for each bootstrap sample $(X_{1,j}, \ldots, X_{20,j})$, compute

$$\hat{\theta}_j^* = \left(\frac{1}{20} \sum_{i=1}^{20} X_{i,j}^* \right)^2,$$

where $j = 1, \ldots, B$.

```
set.seed(232)
B = 500;
bs = replicate(B, rnorm(n, xbar, xvar)); dim(bs)
ThetaHatStar = colMeans(bs)^2; hist(ThetaHatStar)
```

(d) Compute an estimate of the bias:

\begin{align*} \widehat{B}(\widehat{\theta}) \approx \frac{1}{B}\sum^B_{j=1}\widehat{\theta}^_j - *bar{x}*^2. *end{align*}

Compare this to the exact bias using the formula in problem A.1.

```
biasEst = mean(ThetaHatStar) - xbar^2;
bias = 1/n;
cat("The bootstrap estimate of the bias is ", biasEst,". The true bias is ",bias,".")
```

(e) Compute an estimate of the variance:

\begin{align*} \widehat{\text{Var}}(\widehat{\theta}) \approx \frac{1}{B-1}\sum^B_{j=1}\big(\widehat{\theta}^_j - *bar{\theta}* *big)^2, *end{align*} where \begin{align*} \bar{\theta} = \frac{1}{B}\sum^B_{j=1}\widehat{\theta}^_j . *end{align*} Compare this to the exact variance using the formula in problem A.1.

```
vEst = 1/(B-1)*sum((ThetaHatStar - mean(ThetaHatStar))^2);
v = 4*4/n+2/n^2
cat("The bootstrap estimate of the variance of our estimator is ", vEst,". The true variance of our e
```

(f) True or False: For a fixed sample size $n = 20$, as B increases, $\widehat{Bias}(\hat{\theta})$ will approach $Bias(\hat{\theta})$. That is, for a fixed n, the bootstrap estimate of the bias will approach the true bias as the number of bootstrap samples, B increases. You might consider running a simulation to decide!

False. As B increases, the boostrap bias approaches the "plug in" bias estimate, S^2/n, not the true bias σ^2/n.

Problem B.3

The "Wisconsin Card Sorting Test" is widely used by psychiatrists, neurologists, and neurophycologists with patients who have a brain injury. Patients with any sort of frontal lobe lesion generally do poorly on the test. The data frame WCST contains the test scores from a group of 50 patients from the *Virgen del Camino* Hospital.

(a) Using the code below, load the WCST data and explore whether there is reason to believe that the score data comes from a non-normal distribution. First, create a histogram (use ggplot!) and describe whether the data look normal. Then, use the function shapiro.test() to explore normality. Be sure to explain what this function does--i.e., what's the null and alternative hypothesis--in your answer.

```
wcst = read.table("https://www.colorado.edu/amath/sites/default/files/attached-files/wcst.txt")
#or: install.packages("PASWR", dependencies = TRUE)
#library(PASWR); data(WCST)

library(ggplot2)
head(wcst)

#Normality exploration
ggplot(data = wcst) +
geom_histogram(mapping = aes(x = score),color = "black", fill = "white", binwidth = 5)
shapiro.test(wcst$score)
```

The histogram suggests that the data are not normal. The Shiparo-Wilk test for normality tests H_0 : *The data come from a normal population* against H_1 : *The data don't come from a normal population*. The small p-value suggests that we should reject the null hypothesis.

(b) What assumptions must be made in order to compute a (non-boostrap) 95% confidence interval for the population mean score?

We would assume that $\bar{X}$ follows a normal distribution, as suggested by the central limit theorem. Because $n = 50$, this assumption isn't unreasonable.

(c) Compute the confidence interval from (b).

```
n = length(wcst$score); xbar = mean(wcst$score); s = sd(wcst$score);
e = 1.96*(s/sqrt(n))
ci = c(xbar - e, xbar + e); ci
```

(d) Compute a 95% bootstrap pivot confidence interval for the mean.

```
B = 5000; bs = replicate(B, sample(wcst$score, n, replace = TRUE))
xbarstar = colMeans(bs)
q = as.numeric(quantile(xbarstar, probs = c(0.025, 0.975)))
l = 2*xbar - q[2]
u = 2*xbar - q[1]
ci2 = c(l,u); ci2
```

Problem B.4

The dataset gives the number of births per month in New York city, from January 1946 to December 1959. The data are ordered.

(a) Construct another column in the dataset that labels the month and year for each birth per month record.

```
library(tidyverse)
library(ggplot2)
births = read.table("https://robjhyndman.com/tsdldata/data/nybirths.dat", sep = "\t")
births = data.frame(births)
names(births) = "births"
n = length(births$births)
head(births)
```

```
births$month = rep(1:12, n/12)
births$day = rep(1,n)
births$year = rep(1946:1959, each = 12)
births$date = with(births, sprintf("%d-%02d-%02d", year, month, day))
births = as_tibble(births)
births$date = as.Date(births$date,"%Y-%m-%d")
head(births)
```

(b) Construct a plot of births per month against the month/year column that you created in part (a). Analyze the plot. Do you notice anything interesting?

```
p = ggplot(births, aes(x = date, y = births))
p = p + geom_line()
p = p + scale_x_date(date_breaks = "2 years" , date_labels = "%b-%y")
p
```

(c) Suppose that your boss asked you to use the bootstrap to construct a confidence interval for the average number of births per month in New York city over the time period in the dataset. Write a short response to your boss describing why this confidence interval is not valid for this data.

The bootstrap confidence interval would not be valid because the data are not independent and identically distributed. This is clear in the plot above. In particular, there is a clear trend in the mean of births, which violates the identically distributed assumption. The mean decreases from January 1946 to around January 1949, and then increases through December 1959.